Online Supplementary Materials B

1. C-M 95% CIs for Within-subject Designs

In their seminal article, Loftus and Masson (1994) highlighted the importance of adopting different methods to calculate confidence intervals (CIs) for within-subject and between-subjects designs. Therefore, we calculated within-subject CIs according to Baguley’s (2012) proposal, which other researchers (e.g., Cousineau & O’Brien, 2014) commended as appropriate. Specifically, we followed Baguley’s (2012) recommendation of applying the adjusted Cousineau-Morey method to explore the pattern among means and mean differences, by using the cm.ci function in R (v4.0.2; R Core Team, 2019) to calculate the C-M 95% CIs for vocabulary gain scores. For complete R codes of the cm.ci function in Baguley (2012), see <https://osf.io/6768q/?show=view>

1. *Z* Values for Skewness and Kurtosis

To check data normality, we divided skewness and kurtosis statistics by their standard errors to calculate *z* values (Field, 2018; Hair et al., 2019), and interpreted absolute *z* values larger than 2.58 as indicating significant skewness and kurtosis (*p* < .01) in a small-to-moderate sample (Tabachnick & Fidell, 2018).

1. Repeated-Measures ANOVA

Results showed that the assumption of sphericity was not violated (Mauchly’s test of sphericity, *W* = 0.928, *p* = .082). Following the recommendation of reporting corrected results regardless of Mauchly’s test statistics, we reported the results with the Huynh-Feldt correction given *W* > 0.75 (Field, 2018).

1. Bootstrapped Descriptive Statistics; Bootstrapped Pearson’s Correlations

10,000 samples were used for the calculation (LaFlair et al., 2015; Larson-Hall, 2016).

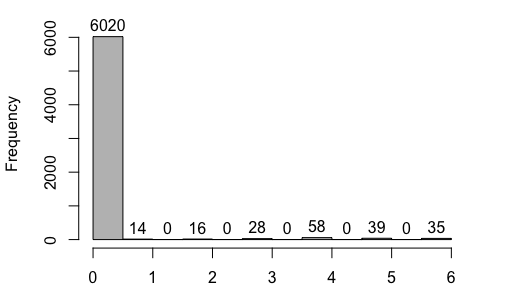
1. Binomial Distribution

For the 0 or 1 scores used with all recognition and the meaning recall items (Test Formats [3], [4], [5], [7], [8]), a binomial distribution is generally appropriate to describe the dichotomous possibilities (Cunnings & Finlayson, 2015; Gelman & Hill, 2007; Zuur & Ieno, 2016; Zuur et al., 2009).

1. Converted Scores with Excessive Zeros

Regarding the form recall items that used fraction scoring between 0 and 1 (Test Formats [1], [2], [6]), the converted scores in integers contained excessive zeros (6020, 96.94%; see Figure 1), and can be modeled by a hurdle model that consists of a binary part and a positive part (Neelon & O’Malley, 2019; Neelon et al., 2016; Zuur & Ieno, 2016; Zuur et al., 2009). For the binary part of the hurdle model, a Bernoulli distribution is usually used; and for the positive part, when overdispersion (i.e., the variance is larger than the mean) does not exist, a truncated Poisson distribution can be adopted, or in the case of overdispersion, a truncated negative binomial distribution (Zuur & Ieno, 2016; Zuur et al., 2009). Alternatively, the Conway-Maxwell-Poisson distribution has been recommended for its flexibility in coping with both under- and over-dispersion in a Poisson distribution, and its truncated form can be used in the positive part of hurdle models (Sellers & Premeaux, 2021).

To examine the dispersion of the converted scores, we used the glmmTMB function from the glmmTMB package (v1.1.2.3; Brooks et al., 2017) in R (v4.1.2; R Core Team, 2021) to build a Poisson mixed effects model, as well as the check\_overdispersion and the check\_zeroinflation functions in the performance package (v0.8.0; Lüdecke et al., 2021) to calculate dispersion and zero ratios. Results showed over-dispersion (dispersion ratio = 1.311) and zero-inflation (ratio of predicted to observed zeros = 0.94) existed. Therefore, for the form recall items, we built a hurdle mixed effects model by using a truncated Conway-Maxwell-Poisson distribution for the positive part and a Bernoulli distribution for the binary part.



*Figure 1. Histogram of converted item-level vocabulary gain scores for the form recall items (Test Formats [1], [2], [6]).*

1. Deviation Coding for Categorical Predictor

Barr et al. (2013) used both treatment coding and deviation coding for mixed effects modeling in their simulation studies, and recommended that deviation coding was generally preferrable than treatment coding in comparing group means. Therefore, we adopted deviation coding for presentation formats as a categorical variable. Specifically, we designated the horizontal format as the reference level, and compared it with the other two more integrated formats, namely, the vertical and the adjacent formats. As Lee and Kalyuga (2011) focused on the horizontal and the vertical formats, the current coding also allowed us to compare our results with theirs. Following Barr (2019), we assigned the coding scheme below to the three formats (see Table 1).

*Table 1. Deviation Coding for Presentation Formats*

|  |  |  |
| --- | --- | --- |
|  | Vertical vs. Horizontal | Adjacent vs. Horizontal |
| Horizontal |  |  |
| Vertical |  |  |
| Adjacent |  |  |

With this coding scheme, we did some calculations in order to interpret the fixed effects estimates for presentation formats by following Field’s (2018) steps. Specifically, we started by using a simplified equation to describe the relationship between vocabulary gain scores (Gain) and presentation formats (ContrastVH = vertical vs. horizontal; ContrastAH = adjacent vs. horizontal):

Gain = b0 + b1ContrastVH + b2ContrastAH (1)

Then we calculated the average gain score for each presentation format (MH = horizontal; MV = vertical; MA = adjacent) by replacing ContrastVH and ContrastAH in equation (1) with the coding values in Table 1:

MH = b0 b1 b2 (2)

MV = b0 + b1 b2 (3)

MA = b0 b1 + b2 (4)

With equations (2) to (4), we calculated the difference between group means for the pairwise comparisons specified in Table 1:

Vertical vs. Horizontal: MV MH = (b0 + b1 b2) (b0 b1 b2) = b1

Adjacent vs. Horizontal: MA MH = (b0 b1 + b2) (b0 b1 b2) = b2

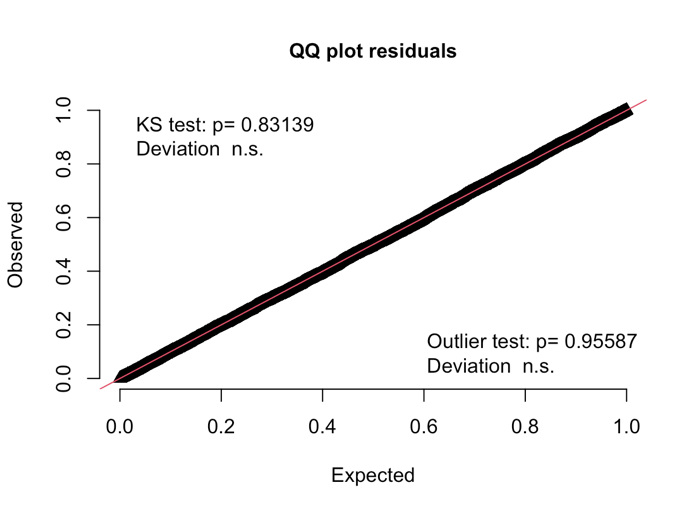
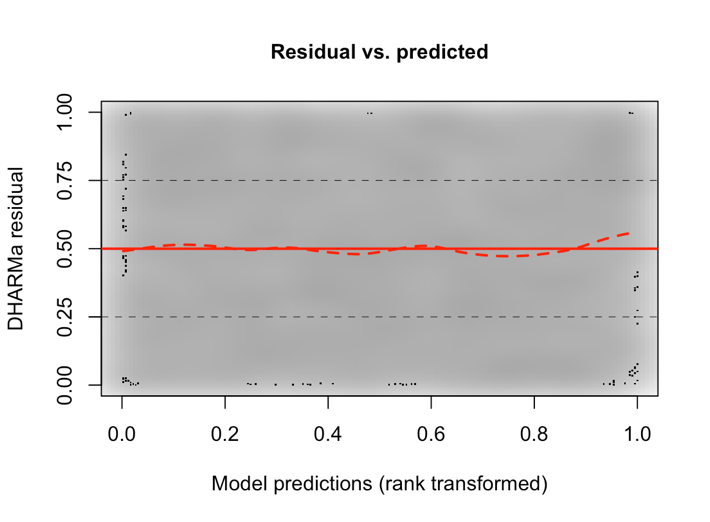
Therefore, with the current coding scheme, we interpreted b1 as the mean difference between vertical and horizontal formats, andb2 as the mean difference between adjacent and horizontal formats.

1. Centering for Continuous Predictors

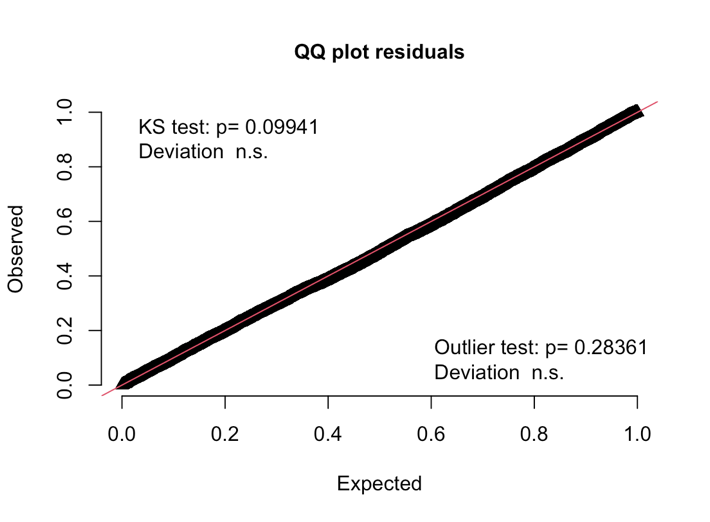
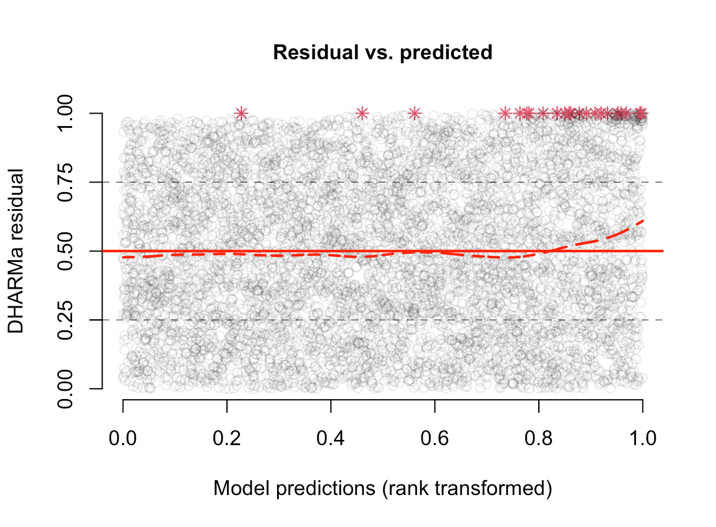
For L2 proficiency scores as a between-subject variable, grand mean centering was conducted with each participant’s score minus all participants’ average score (Brauer & Curtin, 2018). For preference ratings as a within-subject variable, group mean centering was conducted with each rating of a participant minus that participant’s average rating for three formats (Brauer & Curtin, 2018).

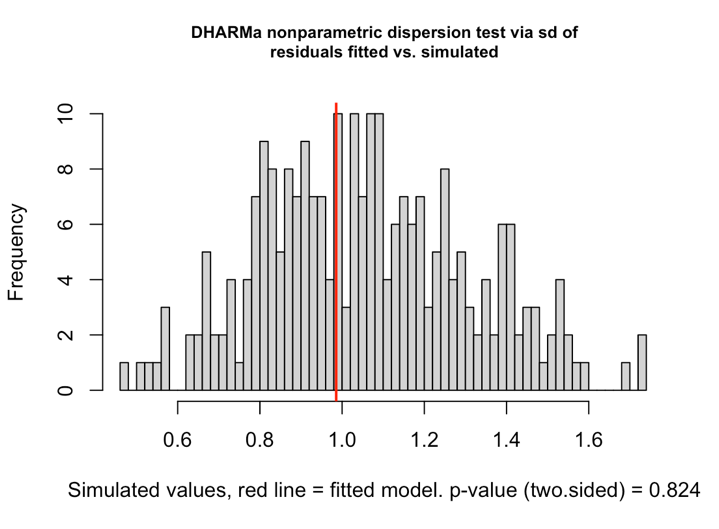
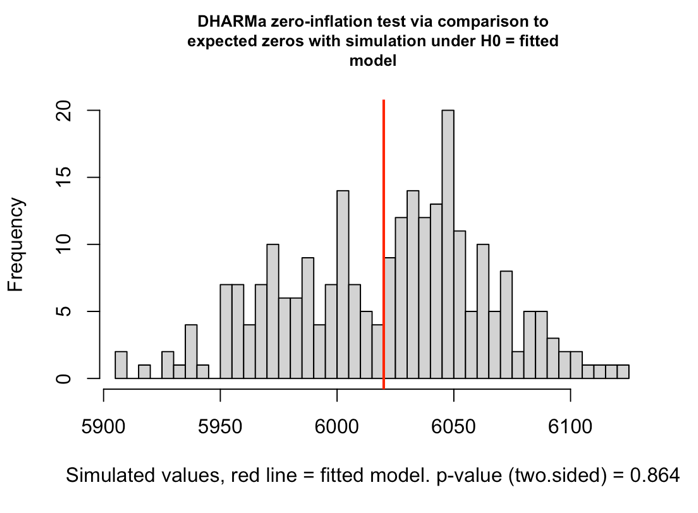
1. Model Diagnostics

Following Gries’ (2021) recommendation, we conducted model diagnostics for the two final mixed effects models, by using the DHARMa package (v0.4.5; Hartig, 2022) in R (v4.1.2; R Core Team, 2021). Specifically, we used the plotQQunif and plotResidual functions to check the residual patterns of the mixed logit model. The residual plots (see Figure 2) did not flag serious issues. For the hurdle mixed effects model, in addition to creating residual plots with the plotQQunif and plotResidual functions, we also used the testDispersion and testZeroInflation functions to check whether over-dispersion or zero-inflation existed in the model. The residual plots (see Figure 3) do not show serious issues except some outliers. The simulation plots (see Figure 3) do not identify over-dispersion or zero-inflation issues in the model. Overall, these model diagnostics indicate the lack of serious problems in the two mixed effects models.

*Figure 2. Residual plots for the mixed logit model.*

*Figure 3. Residual plots and simulation plots for the hurdle mixed effects model.*

References

Baguley, T. (2012). Calculating and graphing within-subject confidence intervals for ANOVA. *Behavior Research Methods*, *44*(1), 158-175.

Barr, D. (2019). Coding categorical predictor variables in factorial designs. <https://talklab.psy.gla.ac.uk/tvw/catpred/>

Barr, D. J., Levy, R., Scheepers, C., & Tily, H. J. (2013). Random effects structure for confirmatory hypothesis testing: Keep it maximal. *Journal of Memory and Language*, *68*(3), 255-278.

Brauer, M., & Curtin, J. J. (2018). Linear mixed-effects models and the analysis of nonindependent data: A unified framework to analyze categorical and continuous independent variables that vary within-subjects and/or within-items. *Psychological Methods*, *23*(3), 389-411.

Brooks, M. E., Kristensen, K., van Benthem, K. J., Magnusson, A., Berg, C. W., Nielsen, A., Skaug, H. J., Maechler, M., & Bolker, B. M. (2017). glmmTMB balances speed and flexibility among packages for zero-inflated generalized linear mixed modeling. *The R Journal*, *9*(2), 378–400.

Cousineau, D., & O’Brien, F. (2014). Error bars in within-subject designs: A comment on Baguley (2012). *Behavior Research Methods*, *46*(4), 1149-1151.

Cunnings, I., & Finlayson, I. (2015). Mixed effects modeling and longitudinal data analysis. In L. Plonsky (Ed.), *Advancing quantitative methods in second language research* (pp.159-181). New York, NY: Routledge.

Field, A. (2018). *Discovering statistics using IBM SPSS Statistics* (5th Ed.). Los Angeles: SAGE.

Gelman, A., & Hill, J. (2007). *Data analysis using regression and multilevel-hierarchical models*. Cambridge: Cambridge University Press.

Gries, S. T. (2021). (Generalized linear) Mixed‐effects modeling: A learner corpus example. *Language Learning*, *71*(3), 757-798.

Hair, J. F., Black, W. C., Babin, B. J., & Anderson, R. E. (2019). *Multivariate data analysis* (8th ed.). Hampshire: Cengage Learning EMEA.

Hartig, F. (2022). DHARMa: Residual diagnostics for hierarchical (multi-level/mixed) regression models (version 0.4.5). Retrieved from https://cran.r-project.org/package=DHARMa

LaFlair, G., Egbert, J., & Plonsky, L. (2015). A practical guide to bootstrapping descriptive statistics, correlations, t tests, and ANOVAs. In L. Plonsky (Ed.), *Advancing quantitative methods in second language research* (pp. 46–77). New York, NY: Routledge.

Larson-Hall, J. (2016). *A guide to doing statistics in second language research using SPSS and R*. London, England: Routledge.

Lee, C. H., & Kalyuga, S. (2011). Effectiveness of different pinyin presentation formats in learning Chinese characters: A cognitive load perspective. *Language Learning*, *61*(4), 1099-1118.

Linck, J. A., & Cunnings, I. (2015). Chapter 8 The utility and application of mixed-effects models in second language research. *Language Learning*, *65*(Suppl. 1), 185-207.

Loftus, G. R., & Masson, M. E. (1994). Using confidence intervals in within-subject designs. *Psychonomic Bulletin & Review*, *1*(4), 476-490.

Lüdecke, D., Ben-Shachar, M., Patil, I., Waggoner, P., & Makowski, D. (2021). performance: An R package for assessment, comparison and testing of statistical models. *Journal of Open Source Software*, *6*(60), 3139.

Meteyard, L., & Davies, R. A. (2020). Best practice guidance for linear mixed-effects models in psychological science. *Journal of Memory and Language*, *112*.

Nakagawa, S., Johnson, P. C., & Schielzeth, H. (2017). The coefficient of determination *R2* and intra-class correlation coefficient from generalized linear mixed-effects models revisited and expanded. *Journal of the Royal Society Interface*, *14*(134), 20170213.

Neelon, B., & O’Malley, A. J. (2019). Two-part models for zero-modified count and semicontinuous data. In A. Levy, S. Goring, C. Gatsonis, B. Sobolev, E. van Ginneken, R. Busse (Eds.), *Health Services Evaluation* (pp. 695-716). Springer.

Neelon, B., O’Malley, A. J., & Smith, V. A. (2016). Modeling zero‐modified count and semicontinuous data in health services research part 1: Background and overview. *Statistics in Medicine*, *35*(27), 5070-5093.

R Core Team. (2019, 2021). R: A language and environment for statistical computing [software]. Vienna, Austria: R Foundation for Statistical Computing.

Sellers, K. F., & Premeaux, B. (2021). Conway–Maxwell–Poisson regression models for dispersed count data. *Wiley Interdisciplinary Reviews: Computational Statistics*, *13*(6).

Tabachnick, B. G., & Fidell, L. S. (2018). *Using multivariate statistics* (8th ed.). New York: Pearson.

Zuur, A., & Ieno, E. N. (2016). *Beginner’s guide to zero-inflated models with R*. Newburgh: Highland Statistics Ltd.

Zuur, A. F., Ieno, E. N, Walker, N. J., Saveliev, A. A., & Smith, G. M. (2009). *Mixed effects models and extensions in ecology with R*. New York, NY: Springer.